What is claimed is:

1. A computer-implemented method of determining discourse structures, the method comprising:

generating a set of one or more discourse parsing decision rules based on a training set; and

determining a discourse structure for an input text segment by applying the generated set of discourse parsing decision rules to the input text segment.

2. The method of claim 1 wherein the training set comprises a plurality of annotated text segments and a plurality of elementary discourse units (EDUs), each annotated text segment being associated with a set of EDUs that collectively represent the annotated text segment.

3. The method of claim 2 wherein the annotated text segments are built manually by human annotators.

4. The method of claim 2 wherein generating the set of discourse parsing decision rules comprises iteratively performing one or more operations on a set of EDUs to incrementally build the annotated text segment associated with the set of EDUs.

5. The method of claim 4 wherein the one or more operations iteratively perform comprise a shift operation and/or one or more reduce operations.

6. The method of claim 5 wherein the reduce operations comprise one or more of the following six operations: reduce-ns, reduce-sn, reduce-nn, reduce-below-ns, reduce-below-sn, reduce-below-nn.

7. The method of claim 5 wherein the six reduce operations and the shift operation are sufficient to derive the discourse tree of any input text segment.

8. The method of claim 1 wherein determining a discourse structure comprises incrementally building a discourse tree for the input text segment.

9. The method of claim 8 wherein incrementally building a discourse tree for the input text segment comprises selectively combining elementary discourse trees (EDTs) into larger discourse tree units.

10. The method of claim 8 wherein incrementally building a discourse tree for the input text segment comprises performing operations on a stack and an input list of elementary discourse

trees (EDTs), one EDT for each elementary discourse unit (EDU) in a set of EDUs corresponding to the input text segment.

11. The method of claim 10 further comprising, prior to determining the discourse structure for the input text segment, segmenting the input text segment into EDUs and inserting the EDUs into the input list.

12. The method of claim 1 wherein determining the discourse structure for the input text segment further comprises:

segmenting the input text segment into elementary discourse units (EDUs);

incrementally building a discourse tree for the input text segment by performing operations on the EDUs to selectively combine the EDUs into larger discourse tree units; and

repeating the incremental building of the discourse tree until all of the EDUs have been combined.

13. The method of claim 12 wherein segmenting the input text segment into EDUs is performed by applying a set of automatically learned discourse segmenting decision rules to the input text segment.

14. The method of claim 13 further comprising generating the set of discourse segmenting decision rules by analyzing a training set.

15. The method of claim 1 wherein the input text segment comprises a clause, a sentence, a paragraph or a treatise.

16. A computer-implemented text parsing method comprising:

generating a set of one or more discourse segmenting decision rules based on a training set; and

determining boundaries in an input text segment by applying the generated set of discourse segmenting decision rules to the input text segment.

17. The method of claim 16 wherein determining boundaries comprises examining each lexeme in the input text segment in order.

18. The method of claim 17 further comprising assigning, for each lexeme, one of the following designations: sentence-break, EDU-break, start-parenthetical, end-parenthetical, and none.

19. The method of claim 17 wherein examining each lexeme in the input text segment comprises associating features with the lexeme based on surrounding context.

20. The method of claim 16 wherein determining boundaries in the input text segment comprises recognizing sentence boundaries, elementary discourse unit (EDU) boundaries, parenthetical starts, and parenthetical ends.

21. A computer-implemented method of generating discourse trees, the method comprising:

segmenting an input text segment into elementary discourse units (EDUs); and

incrementally building a discourse tree for the input text segment by performing operations on the EDUs to selectively combine the EDUs into larger discourse tree units.

22. The method of claim 21 further comprising repeating the incremental building of the discourse tree until all of the EDUs have been combined into a single discourse tree.

23. The method of claim 21 wherein the incremental building of the discourse tree is based on predetermined decision rules.

24. The method of claim 23 wherein the predetermined decision rules comprise automatically learned decision rules.

25. The method of claim 23 further comprising generating the predetermined decisions rules by analyzing a training set of annotated discourse trees.

26. The method of claim 21 wherein the operations performed on the EDUs comprise one or more of the following: shift, reduce-ns, reduce-sn, reduce-nn, reduce-below-ns, reduce-below-sn, reduce-below-nn.

27. A discourse parsing system comprising:

a plurality of automatically learned decision rules;

an input list comprising a plurality of elementary discourse trees (EDTs), each EDT corresponding to an elementary discourse unit (EDU) of an input text segment;

a stack for holding discourse tree segments while a discourse tree for the input text segment is being built; and

a plurality of operators for incrementally building the discourse tree for the input text segment by selectively combining the EDTs into a discourse tree segment according to the plurality of decision rules and moving the discourse tree segment onto the stack.

28. The system of claim 27 further comprising a discourse segmenter for partitioning the input text segment into EDUs and inserting the EDUs into the input list.

29. A computer-implemented method comprising determining a discourse structure for an input text segment by applying a set of automatically learned discourse parsing decision rules to an input text segment.

30. A computer-implemented summarization method comprising:

generating a set of one or more summarization decision rules based on a training set; and

compressing a tree structure by applying the generated set of summarization decision rules to the tree structure.

31. The method of claim 30 wherein the tree structure comprises a discourse tree.

32. The method of claim 30 wherein the tree structure comprises a syntactic tree.

33. The method of claim 30 further comprising generating the tree structure to be compressed by parsing an input text segment.

34. The method of claim 33 wherein the input text segment comprises a clause, a sentence, a paragraph, or a treatise.

35. The method of claim 30 further comprising converting the compressed tree structure into a summarized text segment.

36. The method of claim 35 wherein the summarized text segment is grammatical and coherent.

37. The method of claim 35 wherein the summarized text segment includes sentences not present in a text segment from which the pre-compressed tree structure was generated.

38. The method of claim 30 wherein applying the generated set of summarization decision rules comprises performing a sequence of modification operations on the tree structure.

39. The method of claim 38 wherein the sequence of modification operations comprises one or more of the following: a shift operation, a reduce operation, and a drop operation.

40. The method of claim 39 wherein the reduce operation combines a plurality of trees into a larger tree.

41. The method of claim 39 wherein the drop operation deletes constituents from the tree structure.

42. The method of claim 30 wherein the training set comprises pre-generated long/short tree pairs.

43. The method of claim 42 wherein generating the set of summarization decision rules comprises iteratively performing one or more tree modification operations on a long tree until the paired short tree is realized.

44. The method of claim 43 wherein a plurality of long/short tree pairs are processed to generate a plurality of learning cases.

45. The method of claim 44 wherein generating the set of decision rules comprises applying a learning algorithm to the plurality of learning cases.

46. The method of claim 44 further comprising associating one or more features with each of the learning cases to reflect context.

47. A computer-implemented summarization method comprising:

generating a parse tree for an input text segment; and

iteratively reducing the generated parse tree by selectively eliminating portions of the parse tree.

48. The method of claim 47 wherein the generated parse tree comprises a discourse tree.

49. The method of claim 47 wherein the generated parse tree comprises a syntactic tree.

50. The method of claim 47 wherein the iterative reduction of the parse tree is performed based on a plurality of learned decision rules.

51. The method of claim 47 wherein iteratively reducing the parse tree comprises performing tree modification operations on the parse tree.

52. The method of claim 51 wherein the tree modification operations comprise one or more of the following: a shift operation, a reduce operation, and a drop operation.

53. The method of claim 52 wherein the reduce operation combines a plurality of trees into a larger tree.

54. The method of claim 52 wherein the drop operation deletes constituents from the tree structure.

55. A computer-implemented summarization method comprising:

parsing an input text segment to generate a parse tree for the input segment;

generating a plurality of potential solutions;

applying a statistical model to determine a probability of correctness for each of potential solution;

extracting one or more high-probability solutions based on the solutions' respective determined probabilities of correctness.

56. The method of claim 55 wherein the generated parse tree comprises a discourse tree.

57. The method of claim 55 wherein the generated parse tree comprises a syntactic tree.

58. The method of claim 55 wherein applying a statistical model comprises using a stochastic channel model algorithm.

59. The method of claim 58 wherein using a stochastic channel model algorithm comprises performing minimal operations on a small tree to create a larger tree.

60. The method of claim 58 wherein using a stochastic channel model algorithm comprises probabilistically choosing an expansion template.

61. The method of claim 55 wherein generating a plurality of potential solutions comprises identifying a forest of potential compressions for the parse tree.

62. The method of claim 61 wherein the generated parse tree has one or more nodes, each node having N children (wherein N is an integer), and wherein identifying a forest of potential compressions comprises:

generating $2^N - 1$ new nodes, one node for each non-empty subset of the children; and

packing the newly generated nodes into a whole.

63. The method of claim 61 wherein the generated parse tree has one or more nodes, and wherein identifying a forest of potential compressions comprises assigning an expansion-template probability to each node in the forest.

64. The method of claim 55 wherein extracting one or more high-probability solutions comprises selecting one or more trees based on a combination of each tree's word-bigram and expansion-template score.

65. The method of claim 64 wherein selecting one or more trees comprises selecting a list of trees, one for each possible compression length.

66. The method of claim 55 further comprising normalizing each potential solution based on compression length.

67. The method of claim 55 further comprising, for each potential solution, dividing a log-probability of correctness for the solution by a length of compression for the solution.